Information & Management 52 (2015) 850–858

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

EXPRS: An extended pagerank method for product feature extraction from online consumer reviews

Zhijun Yan [a], Meiming Xing [a], Dongsong Zhang [a,b,*], Baizhang Ma [a]

[a] School of Management & Economics, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing 100081, China
[b] Department of Information Systems, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Received 27 August 2014
Received in revised form 23 January 2015
Accepted 14 February 2015
Available online 7 March 2015

Keywords:
Online product reviews
Feature extraction
Extended PageRank algorithm
Synonym expansion
Social media analytics

ABSTRACT

Online consumer product reviews are a main source for consumers to obtain product information and reduce product uncertainty before making a purchase decision. However, the great volume of product reviews makes it tedious and ineffective for consumers to peruse individual reviews one by one and search for comments on specific product features of their interest. This study proposes a novel method called EXPRS that integrates an extended PageRank algorithm, synonym expansion, and implicit feature inference to extract product features automatically. The empirical evaluation using consumer reviews on three different products shows that EXPRS is more effective than two baseline methods.

1. Introduction

The rapid advance of Internet technology has been greatly propelling e-commerce. iResearch reports that the total online sales in China reached 0.3 trillion dollars in 2013 [21]. Many e-commerce websites and social media platforms, such as Amazon.com, Taobao.com, Eopinion.com, Yelp.com, and Dianping.com, provide an online information sharing channel that allows customers to post their comments on the products that they have bought or consumed and to browse such product reviews of other consumers in order to acquire more knowledge about product quality [7,45]. Research has shown that online consumer product reviews not only have significant impact on consumers' online purchase decisions [9,33], but also are helpful for manufacturers to improve product design and quality and for online retailers to improve their services [2,18,30].

There are two key problems that hinder consumers from fully taking advantage of online consumer product reviews. First, the explosive growth of online product reviews makes it time-consuming and ineffective for a consumer to navigate individual reviews one by one in order to obtain useful information about a product, causing an information overload problem [25]. Previous studies have shown that information overload negatively influences purchase decisions considerably and may lead to fewer consumer purchases [5,8,15]. Therefore, how to deal with such an information overload problem becomes increasingly critical. Although text mining techniques have been applied to online review analysis, they are mainly used for identifying general trends or certain patterns of online consumer reviews instead of coping with information overload.

Second, in reality, individual consumers may have preferences for different product features. For example, some consumers may consider appearance and weight as the most important factors while purchasing a mobile phone, while others may be mainly concerned about its battery life or functions. As a result, reviews that are rated as 'helpful' by some consumers may not really be helpful to others if those reviews do not comment on specific product features of their interest. Most of current platforms of online consumer product reviews do not take needs and preferences of individual consumers into consideration. Few online review platforms organize and present reviews in a personalized, product-feature oriented manner. As a result, it is practically impossible for a consumer to wade through hundreds or even thousands of reviews in order to identify reviews that have commented on specific product features.

* Corresponding author at: Department of Information Systems, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA. Tel.: +1 4104552851.
E-mail addresses: yanzhijun@bit.edu.cn (Z. Yan), minghaizai@163.com (M. Xing), zhangd@umbc.edu (D. Zhang), 295259083@qq.com (B. Ma).

http://dx.doi.org/10.1016/j.im.2015.02.002
0378-7206/

Product feature extraction from online consumer reviews is a key to addressing the above two problems. As a fundamental and important step of consumer review analysis, feature extraction is aimed to identify product features commented in reviews automatically [11,36]. Existing approaches to feature extraction have several limitations. First, product features commented in reviews can be classified into two types: explicit features and implicit features [19]. Explicit features refer to the features mentioned directly in a review. Implicit features are those that are implicitly commented but do not appear in a review. Extraction of implicit features has been largely ignored in previous studies. Second, a product feature may be modified by several different sentiment terms, and one sentiment term can be used to modify different product features. Based on this phenomenon, we treat a term pair of a product feature and its sentiment term as a network node, and relations between different candidate features and sentiment terms can be treated as network edges. Such a network may be helpful for identifying product features. Although a few approaches to product feature extraction from online consumer reviews have explored the incorporation of such relations [20,35], none of them is based on a product feature-sentiment term relation network. Third, prior studies did not take synonyms of product features (e.g., the internal storage of a mobile phone, memory, and RAM) into consideration. As a result, some product features commented in reviews may not be properly identified.

Aiming for improving the accuracy of product feature extraction from online consumer reviews, this paper proposes a novel integrative approach by combining an extended PageRank algorithm that leverages relationships between product features and sentiment terms, feature expansion by adding relevant synonyms, and implicit feature inference. Specifically, the proposed method pre-processes online consumer reviews using a lexical analysis tool, then constructs a network based on term pairs of candidate product features and sentiment words. Next, an extended PageRank algorithm called NodeRank will be applied to rank all term pairs and derive a candidate feature set, which will be expanded by using a synonym lexicon and implicit feature inference to generate a final product feature set.

This study contributes to the literature in three important aspects. First, we develop an extended PageRank algorithm, namely NodeRank, to find possible product features discussed in online consumer reviews by ranking candidate terms. The evaluation shows that this approach significantly outperforms two current feature extraction methods. Second, consumers often use different terms to represent the same product features. They may also describe product features in an informal language, jargons, or even concocted terms [46]. The product features implicitly referred to by those terms are difficult to be identified by existing methods. To address this limitation, the proposed approach infers implicit product features based on the importance of feature-sentiment term pairs. Third, our proposed approach expands an initial feature set by incorporating synonyms of candidate feature terms based on a synonym lexicon. Results of empirical evaluation show the positive effect of this feature expansion on the performance of product feature extraction. The proposed feature extraction method enables personalized, feature-oriented summarization and presentation of online consumer reviews, as some online retailers such as Taobao.com and jd.com have already started to provide. It can significantly alleviate the information overload problem and fasten consumers' purchase decisions. Product manufacturers can also get valuable feedback from consumers on individual product features.

The rest of the paper is organized as follows. Section 2 will introduce the related research on product feature extraction. Then, the proposed product feature extraction method will be presented in Section 3, followed by the description of evaluation and results in Section 4. Finally, we discuss major research findings and practical implications of this study in Section 5.

2. Related work

Current approaches to product feature extraction from online consumer reviews can be classified into three categories: lexicon based, dependency relation based, and machine learning based approaches. We will introduce these three approaches in the rest of this section.

2.1. Lexicon based feature extraction approaches

This type of approach extracts product features based on a lexicon that includes manually predefined product-dependent features. For example, Poria et al. [34] proposed a rule-based approach that exploited common-sense knowledge and sentence dependency trees to identify product features. A predefined implicit feature lexicon was employed to extract product features and an opinion lexicon was used to analyze users' feature inclination. Li et al. [27] firstly constructed a lexicon that contained a list of feature words and opinion words and used it to assign a polarity tag to each feature. Then a natural language processing technique and a statistical method are applied to extract feature words and opinion words based on the defined lexicon.

A lexicon-based approach can be easily implemented. However, the effectiveness of this approach highly depends on predefined feature lexicons. With the rapid increase of online products, it is impractical to manually develop and update feature lexicons for every product.

2.2. Dependency relation based feature extraction approaches

A dependency relation based approach extracts product features based on dependency relations among terms appeared in review sentences. Researchers find that there often exist relations between feature terms and sentiment words in review sentences, which may help extract product features [35]. Such approaches firstly analyze term dependency relations in review sentences, then apply some rules and algorithms to extract product features from the identified dependency relations. For example, Qiu et al. [35] suggested that product features and sentiment words always co-exist with some particular dependency relations. They extracted dependency relations from review sentences using a syntactic parser, and applied the DP (Double Propagation) algorithm to extract product features and sentiment words from them. Zhang et al. [48] used the DP algorithm to extract features from term dependency relations between opinion words and features, and applied the HITS (Hyperlink Induced Topic Search) algorithm to rank features based on their relevance and occurrence frequency. The evaluation result showed that the HITS algorithm was effective for ranking candidate features in feature extraction. Huang et al. [20] treated feature extraction as a sequence labeling task and used the CRF (Conditional Random Fields) model to extract product features from dependency relations. They incorporated part-of-speech features and sentence structure features, which were analyzed by a syntactic dependency parser, into the CRF's learning process. Ahmad et al. [4] mainly analyzed dependency relations of short sentences in reviews, in which dependency relations may be simpler, to improve extraction performance.

Approaches to feature extraction based on dependency relations are product independent and can be applied to large-scale review data. However, because of the complexity of a language, it is difficult to identify all dependency relations in reviews. The omissions of some dependency relations will greatly affect